Document Expansion, Query Translation and Language Modeling for Ad-hoc IR

Johannes Leveling[1], Dong Zhou[2], Gareth F. Jones[1], and Vincent Wade[2]

[1] Centre for Next Generation Localisation
School of Computing
Dublin City University, Dublin 9, Ireland
{johannes.leveling, gareth.jones}@computing.dcu.ie
[2] Centre for Next Generation Localisation
Computer Science Department
Trinity College Dublin, Dublin, Ireland
{dong.zhou, vincent.wade}@cs.tcd.ie

Abstract. For the multilingual ad-hoc document retrieval track (TEL) at CLEF, Trinity College Dublin and Dublin City University participated in collaboration. Our retrieval experiments focused on i) document expansion using an entry vocabulary module, ii) query translation with Google translate and a statistical MT system, and iii) a comparison of the retrieval models BM25 and language modeling (LM). The major results are that document expansion did not increase MAP; topic translation using the statistical MT system resulted in about 70% of the mean average precision (MAP) achieved compared to Google translate, and LM performs equally or slightly better than BM25. The bilingual retrieval French and German to English experiments obtained 89% and 90% of the best MAP for monolingual English.

1 Introduction

The TEL (The European Library) task at CLEF is concerned with ad-hoc information retrieval (IR) [1]. Our IR experiments for the ad-hoc IR task at CLEF 2009 aim at investigating several aspects of retrieval: evaluating document expansion (DE) to obtain longer documents for the TEL collection; applying statistical MT [2] for topic translation and comparing it to Google translate, and comparing retrieval by language modeling (LM) [3] with Okapi BM25 [4].

2 Retrieval Experiments

The Lemur toolkit[3] was employed to index and retrieve documents. Two different retrieval models were used: BM25 [4] with default parameters ($b = 1.2$, $k1 = 2.0$, $k3 = 7$) and LM with Jelinek-Mercer smoothing [3]. TEL documents follow the Dublin Core metadata standard and contain multiple fields including title,

[3] http://www.lemurproject.org/

contributors, language, and subject terms. For different experiments, the text of different document fields was extracted and processed to produce a single flat index. Prior to indexing the documents, their contents were preprocessed with the Snowball stemmer[4] and stopwords were removed. (see [5] for a more detailed description of indexed fields and document preprocessing).

For most runs, pseudo-relevance feedback was applied for query expansion (QE): the top ten ranked documents and 30 terms were used for BM25 and the top five documents and 20 added terms for LM. A variant of query expansion using information from an external resource was also explored for bilingual retrieval (QE2). The top 10 results for the query in the source language were identified and translated with Google translate. Highly co-occurring terms were extracted for query expansion, using the mutual information to calculate co-occurrence and select the highest score for target translation. For the bilingual retrieval experiments, topics were translated using either Google translate (GT)[5] or a statistical machine translation system (MT) [2].

3 Document Preprocessing

The main idea for document expansion was to train a classifier on documents containing a Dewey Decimal Code (DDC) to obtain classification codes for all documents. All classification codes are then replaced with their natural language description, which is added to the document before indexing. The natural language descriptions are available in English only and originate from the OCLC web site[6]. The complete natural languages descriptions for DDC contain 1110 entries of which 933 were actually used in the document collection.

We trained an EVM (Entry Vocabulary Module, [6]) on all documents containing a DDC and applied it to select the top-ranked DDC. Documents with a DDC are expanded before indexing by replacing the code with its natural language description; documents without a DDC are first classified using the EVM and then processed as described above.

4 Results and Conclusions

Results for the ad-hoc IR experiments are shown in Table 1. Some experiments achieved a performance among the top five participants at the TEL track at CLEF 2009, i.e. run DEEN1[7] was 4th in bilingual English (0.3333 MAP), run DE3 was 4th in monolingual German (0.2686 MAP), and run EN3 was 5th in monolingual English (0.3696 MAP).

In all cases, runs with blind relevance feedback to expand queries yield a higher MAP compared to the corresponding runs without blind feedback. The

[4] `http://snowball.tartarus.org/`
[5] `http://translate.google.com/`
[6] `http://www.oclc.org/dewey/`
[7] The prefix TCDDCU has been omitted from the run labels for brevity.

query expansion variant based on external information from web pages found by Google web search did not show the expected results as it degraded the performance (DEEN3 vs. DEEN1).

For the bilingual runs with target language English, 89.9% and 90.1% of the MAP for the best monolingual English runs was achieved for French and German, respectively. Using the MaTrEx system for topic translation achieves a MAP of 70.1% in comparison to topic translation by Google translate (FREN2 vs. FREN1).

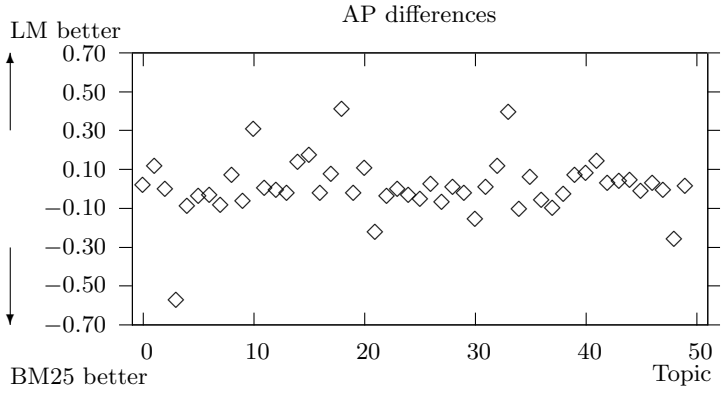

Fig. 1. Differences in AP for English BM25 and LM experiments.

To investigate differences in results for the retrieval models BM25 and LM for monolingual IR, we compared the average precision (AP) for the best runs in English, French, and German (run EN3 vs. EN1F, DE3 vs. DE1F, and FR3 vs. FR1F). A comparison of the English runs EN3 and EN1F is shown in Figure 1. While there seem to be only small changes in performance for the different languages and retrieval models, there is also a small number of topics for each language where the IR models seem to behave very differently. For example for topic 12, LM yields a higher AP compared to BM25 for French; for German, the opposite effect can be observed for this topic. In computing the AP differences, we found that LM returns a higher AP than BM25 for 23 English topics, a lower AP for 26 topics, and the same AP for one topic. For French (German), LM yields a higher AP than BM25 for 29 (23) topics and a lower AP for 21 (27) topics. On average, LM improved precision of slightly less topics compared to BM25, but it resulted in a higher MAP. In conclusion, these IR models seem to return results with similar AP values, but can also behave very differently for certain topics. Further research is required to determine if the best retrieval model for a topic in a given language can be selected automatically or how retrieval results can best be combined.

Table 1. Results for monolingual and bilingual IR experiments for the ad-hoc task.

Run ID	source	target	description	MAP	GMAP	P@10
EN1F	EN	EN	BM25, subset, QE	0.3640	0.1926	0.5080
EN2F	EN	EN	BM25, subset, QE, DE	0.3426	0.1869	0.4980
EN3	EN	EN	LM, subset, QE	0.3696	0.2414	0.5060
EN4	EN	EN	LM, all, QE	0.3688	0.2675	0.5200
FR1	FR	FR	BM25, subset	0.1783	0.0982	0.3340
FR1F	FR	FR	BM25, subset, QE	0.1831	0.0919	0.3420
FR3	FR	FR	LM, subset, QE	0.1758	0.0434	0.2327
FR4	FR	FR	LM, all, QE	0.1749	0.0417	0.2224
DE1	DE	DE	BM25, subset	0.2329	0.1221	0.3540
DE1F	DE	DE	BM25, subset, QE	0.2561	0.1137	0.3580
DE3	DE	DE	LM, subset, QE	0.2686	0.1291	0.3840
DE4	DE	DE	LM, all, QE	0.2439	0.1258	0.3460
DEEN1	DE	EN	LM, GT, subset, QE	0.3333	0.1981	0.4420
DEEN3	DE	EN	LM, GT+QE, subset, QE2	0.2947	0.1351	0.3900
FREN1F	FR	EN	BM25, GT, subset, QE	0.3323	0.1761	0.4820
FREN2	FR	EN	BM25, MT subset,	0.2072	0.0533	0.3800
FREN2F	FR	EN	BM25, MT, subset, QE	0.2551	0.0497	0.3920

References

1. Agirre, E., Di Nunzio, G.M., Ferro, N., Mandl, T., Peters, C.: CLEF 2008: Ad hoc track overview. In: Working Notes for the CLEF 2008 Workshop, 17-19 September, Aarhus, Denmark. (2008)
2. Du, J., He, Y., Penkale, S., Way, A.: MaTrEx: the DCU MT system for WMT 2009. In: Proceedings of the Fourth Workshop on Statistical Machine Translation, Athens, Greece (2009) 95–99
3. Brown, P.F., Cocke, J., Della Pietra, S.A., Della Pietra, V.J., Jelinek, F., Lafferty, J.D., Mercer, R.L., Roossin, P.S.: A statistical approach to machine translation. Computational Linguistics **16**(2) (1990) 79–85
4. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.: Okapi at TREC-3. In: Proceedings of the Third Text REtrieval Conference, Gaithersburg, USA (1994)
5. Leveling, J., Zhou, D., Jones, G., Wade, V.: TCD-DCU at TEL@CLEF 2009: Document expansion, query translation, and language modeling. In: Working Notes of the CLEF 2009 Workshop, 30 September-2 October, Corfu, Greece. (2009)
6. Gey, F.C., Buckland, M., Chen, A., Larson, R.R.: Entry vocabulary – a technology to enhance digital search. In: Proceedings of the First International Conference on Human Language Technology, San Diego, USA (2001)

5 Acknowledgments

This material is based upon works supported by the Science Foundation Ireland under Grant No. 07/CE/I1142.